アラビア文字・多言語文書の横断検索システム構築：「カラム」記事のコーラン引用部分表示の試み

AUTHOR(S):

ブルドン宮本, ジュリアン; 山本, 博之

CITATION:

ブルドン宮本, ジュリアン ...[et al]. アラビア文字・多言語文書の横断検索システム構築：「カラム」記事のコーラン引用部分表示の試み. CIAS discussion paper No.32：「カラム」の時代 IV --マレー・ムスリムによる言論空間の形成 2013, 32: 9-20

ISSUE DATE:

2013-03

URL:

http://hdl.handle.net/2433/228594

アラビア文字・多言語文書の横断検索システム構築
『カラム』記事のコーラン引用部分表示の試み

ブルドン宮本ジュリアン・山本博之

はじめに

本稿は、アラビア文字文書のデジタル・アーカイブの記事に対し、多言語の文書との間の相互参照を検索して表示する検索システムについて、京都大学地域研究統合情報センター（京大地域研）が所蔵・公開する『カラム』（*Qalam*）のデジタル・アーカイブを例として報告する。

本稿が報告する横断検索システムは、地域研究者が研究のために収集した資料群に対して情報技術を用いた研究支援として開発しているものであり、地域研究者のニーズに即した情報技術の開発という意味において、地域研究と情報学のそれぞれにおける要請や意義を背景として行われている。そこで、本稿ではまず本研究の地域研究的背景と情報学的背景をそれぞれ整理し、その上で、『カラム』を例に、資料を収集してデジタル・アーカイブとして公開する過程と、その発展段階として他の文献との間の横断検索システムの事例を報告する。

1. 地域研究的背景

『カラム』は、アフマド・ルトフィ（Ahmad Lutfi）の編集・発行により1950年から1969年までシンガポールで刊行されていたマレー語の月刊誌である[1]。『カラム』には、島嶼部東南アジア各地のさまざまな執筆者により、同時代の出来事についての報告や当時の政治・社会状況に対する意見など、多種多様な記事が掲載されていた。

(1)マレー語のジャウィ表記とローマ字表記

『カラム』の特徴の１つは、創刊から最終号まで全ての記事がジャウィで書かれていたことにある。ジャウィとはアラビア文字を用いたマレー・インドネシア語[2]の表記法で、かつて島嶼部東南アジアでは多くの文献がジャウィで書かれ、在地のムスリムは日常的にジャウィを読み書きしていた。20世紀に入るとしだいにローマ字に切り替えられていき、1950年代以降はほとんどのマレー語定期刊行物がローマ字で刊行されるようになった。

『カラム』は、そのような状況で1960年代末までジャウィによる刊行を継続した数少ないマレー語雑誌だった。ジャウィで書かれているために読者はムスリムにほぼ限定されていたが、他方で『カラム』は国境を越えて読まれており、島嶼部東南アジアにおけるイスラム知識人の公共の言論圏となっていたと言える。

(2)『カラム』の資料価値

『カラム』の研究資料としての価値は、マレーシアの現代史においてイスラム主義運動に関する記述の「空白期間」である1950年代と60年代の社会状況を理解する上での重要性にある。日本軍占領期が終わると、マラヤ・シンガポールではさまざまな政治結社が結成され、民族主義、社会主義、イスラム主義などさまざまな政治結社が作られたが、社会主義勢力とイスラム主義勢力の大半は1940年代末までに植民地当局によって非合法化され、さらに1950年にシンガポールで起こったナドラ事件を契機にこれらの政治勢力はマラヤ・シンガポールの政治の表舞台から姿を消した。この状況は、1970年代に入って官製のダクワ運動が導入されるまで続いた。このように、1950年と60年代のマラヤ・シンガポールのムスリム社会の社会史

1 アフマド・ルトフィおよび『カラム』については［Yamamoto 2009］を参照。

2 島嶼部東南アジアではマレー語をもとにした言語が用いられており、それはマレーシアではマレー語（またはマレーシア語）、インドネシアではインドネシア語、シンガポールとブルネイではマレー語と呼ばれている。それらの言語は、一部の語彙が異なるが、相互にほぼ意思疎通が可能である。本稿では、これらを総称する場合に「マレー・インドネシア語」と呼び、特にマレーシア・シンガポールのマレー語を指すときには「マレー語」と呼ぶ。

は十分に明らかにされていない。しかし、実際には1956年にシンガポールで結成されたムスリム同胞団をはじめ、この時期もムスリムの社会運動は存在していた。『カラム』はムスリム同胞団の事実上の機関誌であり、この研究上の「空白」を埋める格好の素材を提供している。

(3)デジタル・アーカイブ化による資料の共有化

『カラム』には20年間にわたって5,000以上の記事が掲載されたが、これまで『カラム』を主要資料として使った研究はほとんどなかった。その理由として、『カラム』がいくつかの図書館・文書館に分散して所蔵されていることと、ジャウィで書かれているためにイスラム教の背景がない読者には利用が難しいことが挙げられる。また、『カラム』の執筆陣は島嶼部東南アジアの各地に及び、記事の内容は島嶼部東南アジアの各地に及ぶため、当時の広範な地域についての背景知識がないと記事の文脈を捉えにくいという問題も挙げられる。地域研が構築している『カラム』のデジタル・アーカイブ化は、これらの課題を解決し、『カラム』の利用可能性を上げることを目的として構築されている。

いくつかの図書館・文書館に分散して所蔵されていることについては、『カラム』の誌面をデジタル化してインターネット上で公開することで、各機関・個人に分散して所蔵されている資料を仮想的に統合された資料群として利用することが可能になる。また、誌面のデジタル化は、記事本文のOCR処理およびローマ字翻字の可能性が開け、ジャウィに馴染みがない読者にも利用可能性が広がる。このことはさらに、個々の記事中の地名・人名・団体名を他の資料の情報と結びつけて注釈をつけることで、個々の記事の文章の裏に隠された文脈を理解する助けとすることができる。

(4)『カラム』記事のデジタル化の課題

『カラム』はマレー語で書かれた雑誌であるが、ジャウィすなわちアラビア文字で書かれているためもあり、アラビア語文書であるコーラン(クルアン)の章句はアラビア語のまま引用されている。『カラム』の記事にはコーランやハディースからの引用も多く見られ、イスラム研究の専門家ならばそれらの参照元を自力で探し当てることができるだろうが、イスラム教に関する知識を十分に持たない読者のためには、自動処理によってそれぞれの参照元を示すことで利用可能性を高めることができる。

このように、『カラム』の記事の参照で問題になるのは、同じアラビア文字を使いながらアラビア語とマレー語のように言語が異なる文書間で参照がなされている場合、相互の参照関係をどのように検索して示すことができるかである。本稿は、情報技術を利用してこの問題に対応しようとする試みである。

2. 情報学的背景

本節では、『カラム』を例にとって、デジタル・アーカイブに文脈を与える情報学的背景を整理する。まずコンピュータによるジャウィ文字(アラビア文字)の処理について概観し、デジタル・アーカイブ作成における問題点を述べた上で、この問題を解決するセマンティックアノテーションの考え方を紹介する。

(1)コンピュータによるジャウィ文字の処理

ジャウィの処理について検討する前にアラビア文字の処理についてまとめておく。アラビア文字は、個々の文字は独立して表記されるが、単語の中では前後の文字が連結して文字の形が変形するという特徴がある。このため、ローマ字表記や日本語の漢字・かなと異なり、アラビア文字ではまず文字をどう認識するかが問題となる。

ジャウィ字はアラビア文字と同様にユニコードを用いてコンピュータ上で表現することができる。ただし、マレー・インドネシア語にはアラビア語にない発音がいくつかあるため、アラビア文字に点を加えるなどの方法で表記する工夫を行ってきた。たとえば、「nya」、「nga」、「va」はそれぞれ「ڽ」、「ڠ」、「ۏ」と表記する。

最近ではジャウィによる文書もコンピュータで作成されるようになっており、その場合には検索が比較的簡単にできるが、『カラム』などの歴史的な文書はデジタル版が存在しない。そのため、まず原資料をスキャンして電子データにする必要がある。ただし、スキャンしただけではコンピュータは文字として認識しないため、コンピュータ上で文字として読み取り可能にする必要がある。その方法の1つは、文字を1つ1つ人間が読んでコンピュータで入力することだろう。

アラビア文字の光学文字認識(OCR)もあり[Uren *et al*. 2004][M. Zeki *et al*. 2007]、そのジャウィへの適用も試みられている[Omar *et al*. 2012]。ガニーらは、

図1　ジャウィ文書におけるマレー語とアラビア語の文の混在の例

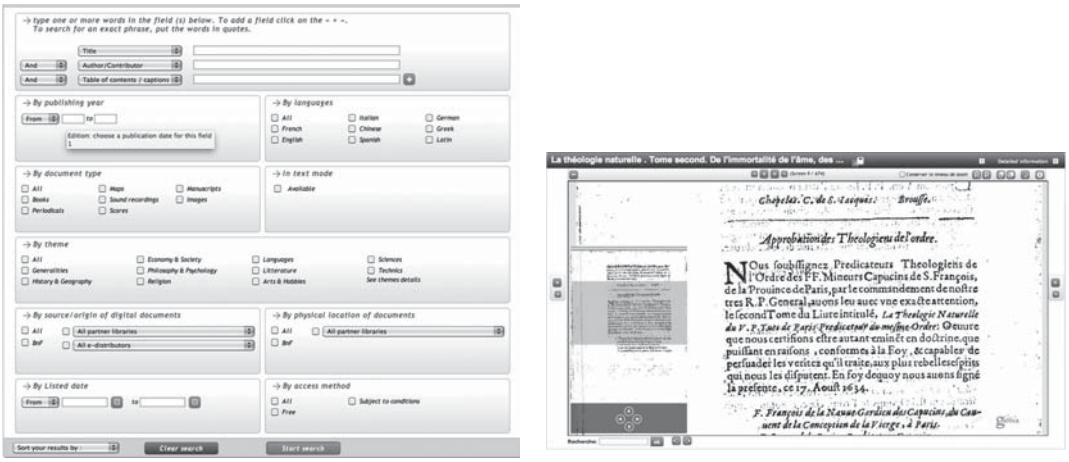

図2　「ガリカ」の検索画面

ステム（幹）に基づくジャウィのローマ字翻字の方法を提案している［Ghani *et al.* 2009］。ただし、ジャウィやアラビア語ではしばしば母音が省略されるため、仮にOCRによって文字として認識されたとしても、適切な母音を補ってローマ字による綴りを得るには辞書を参照する必要がある。また、『カラム』のようなイスラム雑誌では、コーランなどの宗教関係文書からの引用が原語でなされるため、マレー語の文章の中にアラビア語の単語や文が混じるという問題がある。図1は、『カラム』の記事にアラビア語の文が挿入されている例である。上下の部分がマレー語、その間の部分がアラビア語で、図ではわかりやすくするために本文を四角で3つに区切っている。

(2)歴史文書データベース

文献資料をデジタル化する技術は急速に進展している。一般的なカメラの約200倍以上もある1,200メガピクセルの高解像度で原資料を撮影し、文字が書きこまれた羊皮紙の表面も細部までズームして見ることができる死海文書デジタル化プロジェクト（Dead Sea Scrolls digital initiative）［Broshi 2004］に見られるように、デジタル・ライブラリーの構築によって原資料に当たらずとも歴史文書を参照できるようになりつつある。デジタル・アーカイブでは原資料をもとの所有者の手元に置いたまま資料を収集・公開できるため、資料の収奪を防ぎ、また、遠隔地にいる複数の利用者が同時に同一の資料を利用することも可能になる。

ニール・ビーグリー（Neil Beagrie）は、オーストラリア、フランス、オランダ、イギリスのデジタル・ライブラリーを比較検討した［Beagrie 2003］。今日、最も包括的なデジタル・アーカイブの1つはフランス国立図書館（BNF）の電子図書館「ガリカ」（Gallica）だろう。所蔵データは、書籍、地図、手稿、画像、定期刊行物、譜面、音源を含み210万件以上に及ぶ。図2に示したように、ガリカはライブラリ型の検索システムを使っている。すなわち、利用者は記事名や著者名や資料の種類などを入力して対象を絞り込むという方法である。データの一部は本文がデジタル化されており、文献本文の全文検索も可能である。

ジャウィ文書に関しては、現在、北イリノイ大学の

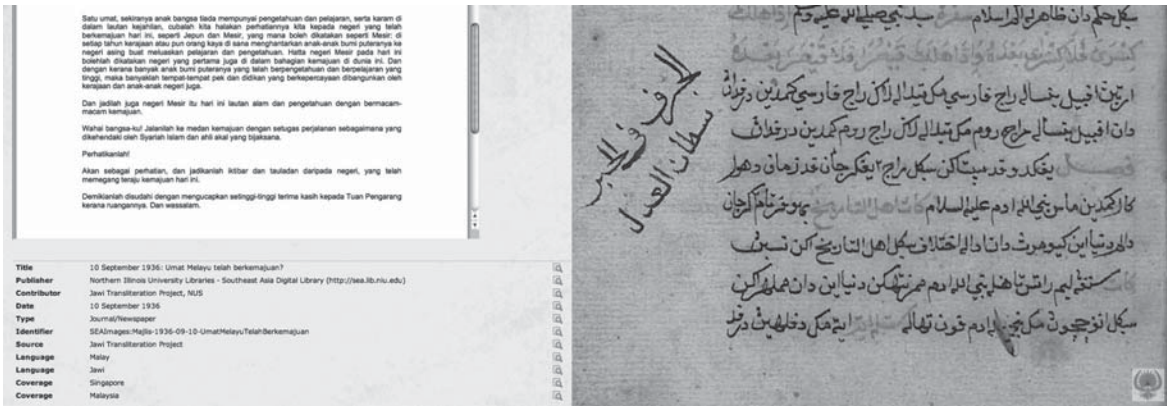

図3　北イリノイ大学ジャウィ文献翻字プロジェクト（左）とマラヤ大学マイマヌスクリプト・データベース（右）

ジャウィ文献翻字プロジェクトとマラヤ大学のマイマヌスクリプト（MyManuskript）データベース・プロジェクトの2つが見られる。

ジャウィ文献のデジタル・アーカイブの研究上の活用の例として、パッテンは、『マジュリス』（Majlis）などのジャウィ雑誌をローマ字翻字し、本文の分析を通じて1930年代前半のマラヤにおける商業主義の浸透の様子を明らかにした［van der Putten 2010］。研究以外の活用の例としては、マレー語文献のデジタル・アーカイブの重要性を教育の見地から論じたザヒダらが、授業以外の場で子どもたちがジャウィに触れる機会を増やし、子どもたちがジャウィについて話をしたり書いたりする機会を増やす意義があると述べている［Zahidah *et al.* 2011］。

一般にデジタル・アーカイブは汎用性と操作性の両立という問題を抱えている。用途を具体的に想定して開発すれば特定の利用者にとって使いやすくなるが、その用途以外の利用者には使いづらくなる。逆に、どの用途にも使えるように設計すると、かえってどの利用者にも使いづらいものになるという問題である。上で紹介したデジタル・アーカイブの多くは、特定の用途を想定しない汎用的なシステムであるライブラリ型の検索システムによってデータが提示されている。ライブラリ型の検索システムでは、利用者は記事名、著者名、文書の種類などによって対象記事を絞り込むため、利用者が資料の全体像を把握していない場合には十分に利用できないことになる。

(3)セマンティックアノテーション（意味的注釈）

利用者が事前に資料の全体像を把握していない状況でデジタル・アーカイブを効果的に活用する方法がいくつか検討されている。テキスト全体をデジタル化して本文を検索対象にすることはその1つであり、ジャウィ文書では本文のローマ字翻字がその一歩に当たる。ただしこの方法では、検索可能な範囲が記事名や執筆者名からテキスト全体に広がることにはなっても、意味による検索ができるようになるわけではない。つまり、「○○氏は過去にどの雑誌に記事を執筆したのか」、「△△に関する記事で最も多く引用されているコーランの章句はどれか」といった質問には答えられない。これらの質問に答えられるシステムを含んだデータベースを構築するには、テキストのデジタル化だけでは不十分であり、各記事に関する情報を含めてデジタル化する必要がある。

ジョン・リー（John K. Lee）とブレンダン・チャランドラ（Brendan Calandra）が示したように、記事に関する情報を含めたデータベース化によって複雑な質問に答えることが可能になる［Lee & Brendan 2004］。リーとチャランドラは、米国憲法に関する2つの異なるウェブサイトを高校生に見せて、意味的注釈が内容の理解を促すかを実験した。一方のウェブサイトには憲法の条文だけ書かれているのに対し、もう一方のウェブサイトにはそれぞれの条文を理解する上で鍵となる言葉に注釈を付したところ、注釈を付したウェブサイトを見た高校生の方が相対的に問題解決の度合いが高くなるという結果が得られた。

それでは、データベース上の記事にどのようにして意味的注釈をつけるのか。特定の文書に関する知識を構築するにはセマンティックアノテーション（意味的注釈）の考え方が利用できる。図4に示されているように、セマンティック・ウェブ（意味的ウェブ）は、それ自体が新しい1つの技術ではなく、知識推論から特定の領域に関する知識までを扱う複数のツールの組み合わせである。

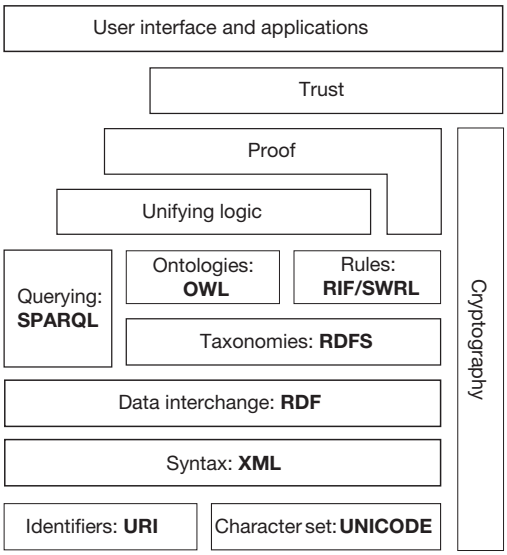

図4　セマンティック･ウェブ(Tim Berners-Leeによる)

イギリスの計算機学者のティム・バーナーズ＝リーらは、人間が読む「文書のウェブ」を機械処理によってデータを発見して利用できる「データのウェブ」に変えるため、セマンティック・ウェブ(semantic web)という考え方を提唱した[Berners-Lee 2001]。この考え方により、複数の情報源から条件に合ったデータを探し出して結び付けることが可能になる。本稿では、この考え方を用いて、『カラム』の記事に見られる他の文献からの参照箇所を機械処理によって表示する方法を提示する。

実際にどのように働くのかを少し丁寧に見てみよう。「Ahmad Lutfi」という単語を例にとる。知識推論を加えなければ、それは11文字のローマ字が繋がったものとしての意味しか持たない。背景知識を持った人であれば、この２つの単語は人物の名前で、その人物は『カラム』の編集者であって、本名はSyed Abdullah bin Abdul Hamid al-Edrusであるといったことがわかるだろう。この人物が『ワルタ・マラヤ』という雑誌の編集にも関わっており、この雑誌は北イリノイ大学の翻字プロジェクトの対象であることを知っている人もいるかもしれない。

ここで逆の例、つまりAhmad Lutfiについての背景知識がない人で、この人物が書いた記事をすべて読みたいと思っている人がいるとしよう。これを自動で調べるには、検索対象となる全ての文書に「著者」という情報が指定されていなければならない。また、この人物が複数の筆名を使っていた場合には筆名の対照表も必要となる。

「Ahmad Lutfi」という検索語を著者と結びつける処理も必要である。ビクトリア・ウレン(Victoria Uren)らが述べたセマンティックアノテーションのシステムが利用できる[Uren *et al.* 2006]。ウレンらは、セマンティックアノテーション・システムが成り立つために必要な条件を挙げている。『カラム』のデジタル・アーカイブに照らしてみれば、意味的注釈はRDF/XMLのような標準的な仕組みで記述される必要があること、記事の文脈を拡張するための知識を複数の利用者が追加できること、文書全体を人の手で意味づけすることは難しいためにある程度まで自動処理がなされていることなどの条件である。これはブディとブレッサンがインドネシアの新聞で試みた固有表現抽出(named entities extraction)によって実現できる[Budi and Bressan 2007]。

3. 地域研究、情報学、ジャウィのローマ字翻字

デジタル・アーカイブ構築の具体的な作業について紹介する前に、ジャウィ文献をローマ字に翻字する意義はどこにあるのかを考えておきたい。

アラビア文字で書かれた文献をローマ字に翻字する意義について、アラブ・中東地域を対象とする研究者にはその意義があまり理解できないかもしれない。なぜなら、アラブ・中東地域の研究者なら、アラビア文字の読み書きを身につけて直接アラビア語で文献を読めばよいと考えるだろうためである。日本研究に置き換えて考えれば、日本で発行された書物を読んで日本のことを知ろうとしたとき、漢字・かなの読み書きを一切勉強せずにローマ字だけで通そうとする人がいたとしたら、その人の日本社会に対する理解度はかなり怪しいという印象を与えることだろう。それと同じで、アラビア文字がどんなに難しそうに見えても、アラブ・中東地域を研究するならアラビア語が読めなければならないという考え方はよく理解できる。

しかし、ジャウィはそれとは事情が異なっている。ジャウィはアラビア文字を改変したマレー・インドネシア語の表記法であり、島嶼部東南アジアではかつて広く用いられ、ムスリムを中心に多くの人がジャウィを読んだり書いたりすることができた。しかし、今日では社会生活の多くの場面でローマ字が用いられており、在地のムスリムでも、特に若い世代では、ジャウィの読み書きが全くできないか、できるとしてもかなりの困難が伴うことが珍しくない。このため、わず

か50年前に書かれたことが今日の読者層には十分に読むことができず、知の継承における深刻な問題があると言える。最近ではマレーシアでムスリムを主な対象に学校教育でジャウィを教える努力がなされるようになっているが、その努力とは別に、ジャウィの読み書きができない人たちにもジャウィで書かれた文書の内容が理解できるような工夫が必要だろう。「アラビア文字が読めるようになればよい」ではなく「アラビア文字が読めない人にも書かれた内容がわかるようなシステムを構築する」という観点からこの問題に取り組むことが本稿の意図である。

このように考えるならば、イスラム教が他宗教と出会って宗教混成社会を作り、聖典の言葉であるアラビア語の単語が現地語であるマレー・インドネシア語の中にそのままの形で使われている東南アジアであるからこそ、アラビア文字・多言語文書の横断検索システムが必要とされると言える。

このことは、京都大学地域研究統合情報センター（京大地域研）が進めている「新しい地域研究」のあり方と重なるところがある。『地域研究』の第12巻第２号（総特集「地域研究方法論」）で示されているように、今日の地域研究は３つの層で捉えることができる（図５）。上から第一層、第二層、第三層と呼ぶことにしよう。一般に地域研究と言えば、特定地域に深くコミットして研究する立場が想像されることが多い。現地語を習得し、長期のフィールドワークによって現地事情に通じた上で研究を行うのは第二層にあたる。これに対して第三層は、そのようにして得られた地域研究の知見を異業種・異分野の専門家に伝わる形で提示する方法を探る地域研究である。そのためにはさまざまな技術や工夫の助けを借りる必要があり、情報技術を用いて現地語文献の概要を把握することも第三層に含まれる。京大地域研は、所属する個々の研究者は第二層の研究を進めているが、その上で、京大地域研全体では地域情報学プロジェクトや「災害対応の地域研究」プロジェクトなどにより第三層の地域研究を進めている。本書で紹介するアラビア文字・多言語文献の横断検索システムが構想された背景の1つにはそのことがある。

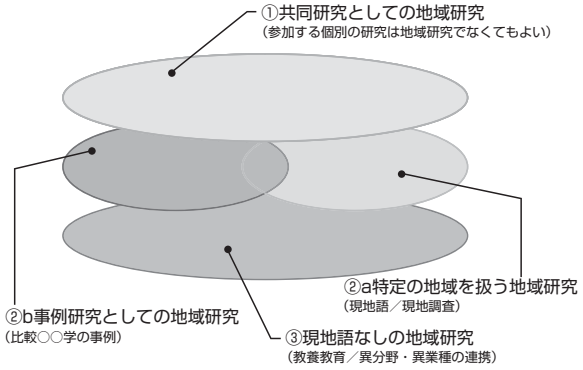

図５　「地域研究」の３つの層

4. イスラム雑誌デジタル・アーカイブの構築

本節では、『カラム』を例にとり、イスラム雑誌のデジタル・アーカイブ構築のための５つのステップを紹介する。すなわち、デジタル化、公開、ローマ字翻字、インデックス作成、注釈作成である。

以下では、一般的な注意事項を含め、デジタル・アーカイブの構築のために必要な手順を具体的に示すことにする。『カラム』はジャウィ文字を用いた紙媒体の雑誌であるため、記事をデジタル・アーカイブ化する上で『カラム』に固有の作業が必要となるものもあるが、ここで示した手順は基本的にジャウィ表記以外の文献のデジタル・アーカイブ化とも共通している。

(1)資料の収集・デジタル化

デジタル・アーカイブの対象となる『カラム』の記事を収集する。最近の刊行物では、紙媒体とともに電子媒体でも提供されていたり、あるいは電子媒体のみで発行されたりしているものがあり、その場合は電子媒体での刊行物を手に入れれば収集とデジタル化が同時に進むことになる。また、従来の資料収集では現物を収集することに重きが置かれ、紙コピーやマイクロフィルムやマイクロフィッシュなどの複写資料は二次的な資料として見なされる傾向もあったが、デジタル・アーカイブ化ではデジタル化された複写資料をデータとするため、必ずしも原資料を入手する必要はないことになる。そのため、資料が管理されている場所でコピー機やスキャナで複写をとったり写真撮影したりする方法をとることもできる（デジタル・アーカイブと別に原資料を収集・所蔵したいかどうかは別の問題である）

『カラム』は紙媒体でのみ刊行された雑誌であるため、現物またはその複写を入手するか、マイクロフィルムなどの形態で資料を入手することになる。『カラム』は、マラヤ大学のザアバ記念図書室に現物が、シンガポール国立大学の図書館にマイクロフィルム２巻分がそれぞれ所蔵されており、両者をあわせるとかなりの巻号がカバーできるが、完全ではない。地域研で

は、この２か所のコレクションを収集した上で、マレーシアの国立図書館や各地の図書館・図書室や個人を尋ねて欠けている巻号を収集し、全体で欠号率が低い『カラム』コレクションとした。このように、刊行から時が経過した刊行物では、一か所の図書館・文書館で全ての巻号が揃っているとは限らず、複数の図書館・文書館や個人収集家の協力を得なければならないこともある。

資料の所在地を特定したら、誌面をスキャンしてデジタル・データを作成する。原資料ではなく複写をもとにデジタル化する場合、「コピーのコピー」となって画質が悪くなるために工夫が必要となる。紙コピーやマイクロフィルムでは印刷の汚れが黒い点々のようになって見えることがあり、同じ形の文字でも点がどこにいくつ打たれているかで違う文字になるジャウィ（アラビア文字）の文献では、それが文字の一部の点なのか印刷の汚れなのかが決定的に重要となる。そのため、例えば、白黒ではなくグレースケールでスキャンした上で汚れの部分を薄くする処理を施したりする。（この方法では、ページ当たりのデータの容量が大きくなることや、ページあたりの作業量が増え、時間とコストがかかることなどの問題もある。）また、特に新聞や雑誌の写真の部分は普通に白黒でコピーすると真っ黒になってしまうため、資料の状態によっては、例えば文字の部分と写真の部分を別の解像度でデジタル化して貼りあわせて１つの資料にするなどの工夫も必要になる。

(2)データの整理と公開

スキャンした誌面は１ページごとの画像データになっている。１つの記事が２ページ以上にわたる場合は記事ごとにまとめ、記事ごとにPDFファイルにする。『カラム』などの一部の雑誌では、記事が複数ページにわたり、最後の部分が１ページに満たない場合、記事の最後の部分は別の記事の下の部分に掲載されることがある。たとえば、ある記事の掲載ページは１ページから４ページまでと33ページというように、最後の部分が飛んでいることがある。

この作業によって記事ごとのPDFファイルがたくさん作られるため、どのファイルがどの記事に当たるかを示す見取り図を作らなければならない。そのためには、PDFファイルに適切な名前を付けることと、それぞれのPDFファイルの記事の情報を記した一覧を作る必要がある。

PDFファイルの名前は、多数のPDFファイルをファイル名で並べたときに探しやすいことを考慮して付ける。新聞・雑誌であれば、例えば原資料の刊行日＋新聞・雑誌名＋掲載頁とし、同じ名前のPDFファイルが複数できる場合にはさらに末尾に記事タイトルや執筆者名（あるいはその最初の１、２語）を添えるなど工夫する。ファイル名を記事名や著者名から始めると、ファイル名で並べたときに刊行日順にならなくなるので注意を要する。また、原資料の刊行日や新聞・雑誌名はフォルダ名にして、フォルダを階層構造にすることでファイル名を短くすることもできるが、その場合には個別のPDFファイルを取り出したときに刊行日や媒体名がわからなくなる可能性があることに注意を要する。

記事の情報の一覧は、記事ごとのタイトル、掲載日、執筆者名、コラム名、PDFファイル名などを記載した一覧表をエクセルファイルで作成する。このエクセルファイルが作れれば、簡易検索エンジンを利用して、先にデジタル化した記事のPDFファイル群を記事データベースとして公開し、記事名や執筆者名による検索が可能になる。PDFファイル群の検索はこのエクセルファイルを通じて行うため、このエクセルファイルの項目をどう作るかは、データベースの検索システムをどう作るかという問題と直結している。その際には、その資料の特徴をうまく引き出すような検索項目の立て方と、他の資料群との横断検索を可能にするような検索項目の立て方という２つの方向性をうまく合致させる必要がある。

このエクセルファイルに項目として挙げられているもののみが検索の対象となるため、検索の対象としたい情報はすべてエクセルファイルに項目を立てておかなければならない。また、他の雑誌から同様に作成した記事データベースとの間でも、エクセルファイルの項目が対応していれば横断検索が可能になる。
他の資料群との横断検索の可能性を考えるのであれば、それが現在扱っている資料群に照らして自明なことであっても、その資料群の性格を示す項目をエクセルファイルに立てることも必要となる。例えば『カラム』の記事データベースでは、全てのデータが『カラム』の記事であることは自明であり、エクセルファイルに「掲載誌名」や「発行国」といった項目を立てることにはほとんど意味がない。ただし、『カラム』の記事データベースを例えばインドネシアで刊行されていた『ワクトゥ』の記事データベースとあわせて検索しようとするならば、「掲載誌名」や「発行国」、さらに「使

用言語」、「使用文字」の情報も必要になるかもしれない。この考え方を進めて、雑誌以外の資料群との横断検索の可能性を想定するならば、「媒体の種類」という項目を立てて、その列の全てのセルに「雑誌」と書いておく必要があるかもしれない。

他の種類の資料群と統合する場合には統合するときになってエクセルファイルに同じデータが入った１列を追加すればよいが、「コラム名」や「執筆者名」などの記事固有の情報は最初に作っておく必要がある。また、エクセルファイルの項目を考える際には、各項目のデータをどの言葉（どの文字）で書くかについても考える必要がある。例えば「発行国」の項目に「Singapore」と「シンガポール」とあったとき、人間の目で見れば同じものだと判断できるが、コンピュータはそれらを別物と判断するため、検索結果が正確でなくなる。国名はどの言語でも一対一で対応するだろうから機械的に対応関係を処理しやすいし、それ以外の項目についても機械翻訳の精度が上がれば多くの面で問題が解消すると思われるが、異なる言語の間で全ての概念を一対一で対応させられるわけではないことを考えると、地域研究のデータベースにおいては、多少手間はかかるが、現地語および想定される利用者の言語のそれぞれについて項目を作っておいた方がよいかもしれない。『カラム』データベースやマレーシア映画データベースでは、エクセルファイルのそれぞれの項目についてマレー語、日本語、英語の３つの列を作っている。

なお、『カラム』データベースでは採用していないが、新聞記事データベースでは、検索対象の項目に「第一段落」を入れるようにしている。これは、インド洋津波の発生直後に新聞記事を大量に収集して情報を整理した際に、記事内容の検索のために本文を入力して検索可能にしようとしたが、記事の量が多かったために全文を入力する余裕がなかった状況で考えられた方法で、新聞の報道記事では第一段落に概要が書かれているため、第一段落だけ入力して検索可能にしておけば収集した記事群の概要を掴みやすいという経験に基づいている。新聞のオンラインでの配信が進めば全文検索が容易になり、その際には新聞記事の検索方法が別の形をとっていくかもしれないが、全文検索が容易でない状況で新聞記事のデータベースを作る場合には第一段落を抜き出す方法が有効だと思われる。

(3)ローマ字翻字

この項目は『カラム』の記事データベースに特徴的な作業である。前項では資料をデジタル化してPDFファイルを作成し、データベースとして公開する段階に至ったが、スキャンした結果をPDFファイルにしたものは画像ファイルであるため、人間の目には文字として認識できても、コンピュータは文字として認識して処理することができない。そのため、それぞれの記事に「記事名」、「執筆者名」などのキーワードを付けたエクセルファイルを作り、それに基づいて検索を行っていた。新聞記事の第一段落を入力するというのも、記事全文が検索の対象となっていないため、なるべく少ない手間で効果的な検索結果が得られる工夫として考えられた方法だった。

この次に考えるべきことは、記事の本文全文を対象とした検索となるだろう。前項でも全文検索について触れたが、新聞・雑誌のオンライン化が進み、あるいは電子書籍の刊行が進めば、今後、刊行物の全文検索が容易になる状況が生じるかもしれない。ただし、過去の出版物やデジタル化されていない資料を対象とする場合には、全文検索を行うためにいくつかの段階が必要となる。いずれも専門性が求められ手間暇がかかる作業であるため、資料の性格やそれをどのように使うかによってとる方法が異なる。

石や木などに刻まれた文字は１つ１つ手作業で写すしかないが、紙媒体の文字資料についてはOCR認識によって文字をテキスト化することができる。OCR認識では精度が問題となるが、ローマ字では十分な精度のものが多くあり、日本語についても実用に足りるものが少なくない。ただし、『カラム』が使っているジャウィでは、コンピュータを用いてジャウィで書かれた文書を機械的にローマ字に翻字するソフトウェアはあるが[Ghani *et al*. 2009]、紙媒体の文書をスキャンした画像をもとにジャウィのOCR認識を十分な精度で行う環境は整っていない。

このため、現状ではジャウィ文字に通じた専門家が記事を読みながら手作業でローマ字翻字している。この部分は自動化できていないために最も手間と時間がかかっているが、翻字の作業と同時に注釈作成なども進めることができる。本稿が掲載される予定のディスカッション・ペーパーは、『カラム』の翻字作業を行っている専門家がそれぞれの専門と関心に応じて『カラム』の内容を分析して研究成果をまとめたものであり、『カラム』ローマ字翻字プロジェクトとしては副産物であると言える。

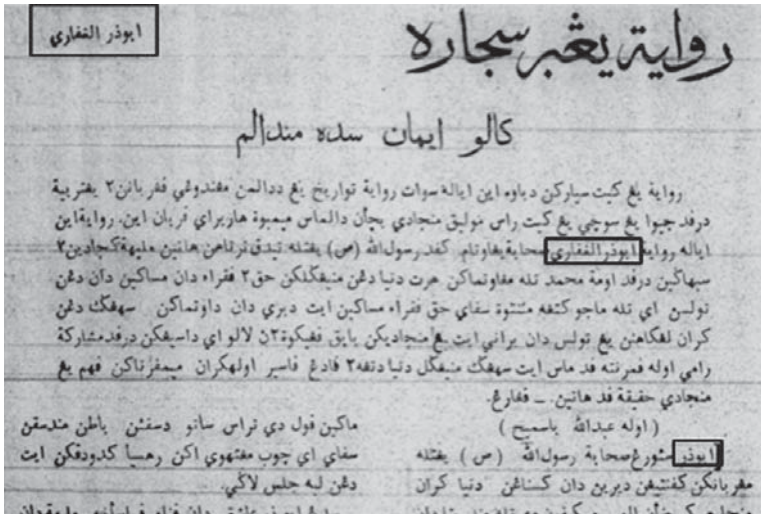

図6　記事参照の例

(4)インデックス作成

記事のローマ字翻字によって記事の全文が検索の対象となる。全文検索では辞書を用いたシステムと形態素解析を行い［Ahmad *et al.* 1996］。これらのアルゴリズムをデータベースと組み合わせる。

(5)参照作成

デジタル化した文書は、文書単位と記事単位の2つのレベルで参照を示すことができる。なお、参照はオントロジーのセットと関連付けする必要がある。

(a)文書単位での参照

［Kakali *et al.* 2012］のように、公開の段階で得られたインデックスは自動的に収集してオントロジーに関連付けることができる。書誌情報の参照だけでなく、文書ごとの参照ではGoogle Scholarや手動のサービス［Bourdon & Shibayama 2012］を利用して他の文献との相互参照を示すことができる。

(b)記事単位での参照

文書単位での参照のほかに、記事の一部に注釈をつけることも可能である。図6は『カラム』の記事から「Abu Dzar Al-Ghifari」という名前を検索したもので、その結果が四角で囲まれて示されている。さらにインターネット上のWikipediaの同名の項目などにリンクを張ることもできる。

5. 横断検索システムの構築 ——『カラム』記事のコーラン章句引用

本節では、アラビア文字文献データベースに他言語の文献からの参照を示す方法として、『カラム』の記事中に引用されているコーランの章句を外部の文書からリンクさせるシステムを紹介する。

『カラム』の記事が書かれているのはジャウィ表記のマレー語だが、記事中でコーランの章句が引用されている部分はコーランに記載されている通りアラビア語で書かれている。同じアラビア文字を使っているが、マレー語の文章の中に一部だけアラビア語の文章が挿入されている。以下で示すのは、『カラム』記事中のコーランの引用部分について、『カラム』とは別の電子版のコーランの章句データベースから該当する記事を検索して示すシステムである。なお、ここで示す検索画面は開発用の仮のものであり、一般公開に当たっては利用者用の使いやすさを考えた検索画面を作る予定である。

(1)データベース

ジャウィで書かれた『カラム』の記事を直接コンピュータで処理できないため、各記事のローマ字版を用いて、それをもとにコーランの該当する章句を探し、さらにそれに対応する『カラム』の記事に結びつけることで、『カラム』の記事とコーランの章句を結びつける。

ローマ字翻字された記事本文にインデックスを付ける。そのため、翻字されたローマ字のPDFファイルからテキストを直接抽出するJavaプログラムを構築した。記事名、執筆者名、PDFファイル名などの基礎的なインデックスはデータベース公開の過程で作成したエクセルファイルの項目を用いた。抽出した内容はPostgreSQLデータベースに格納した［Ahmad *et al.* 1996］。

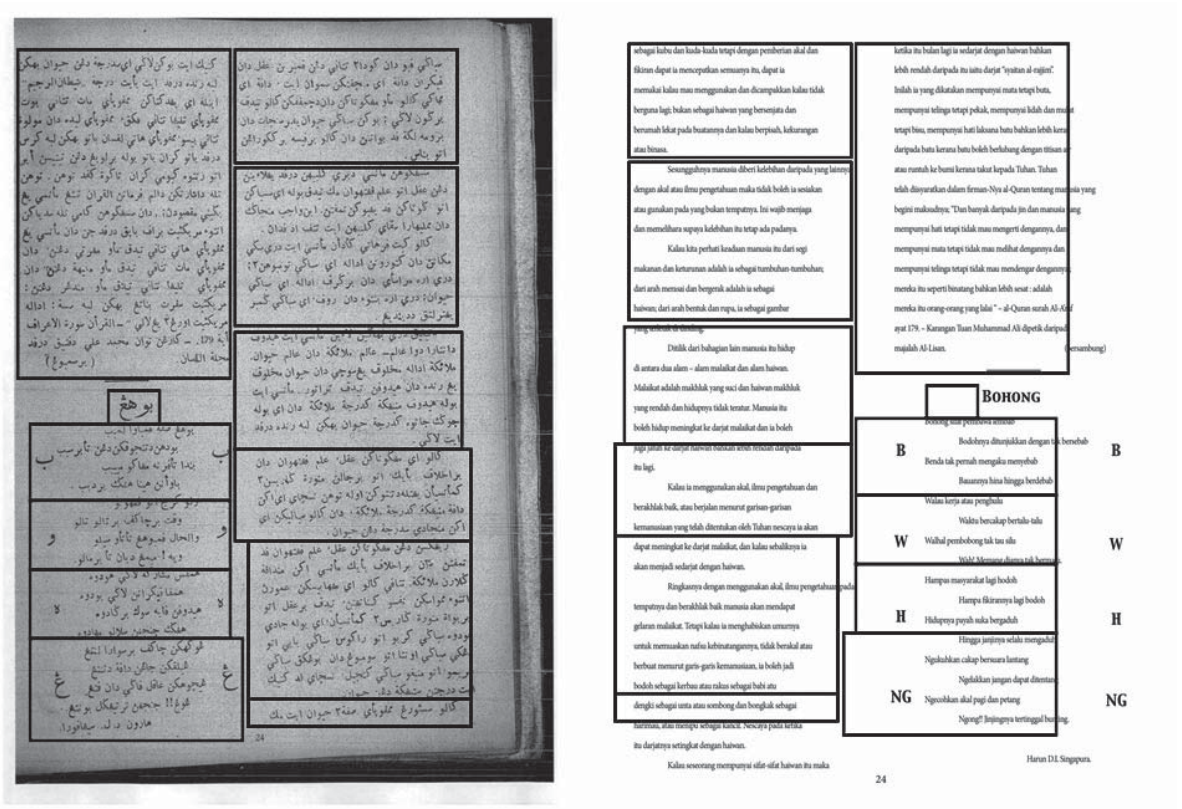

図7　ジャウィとローマ字の領域区切りのずれ

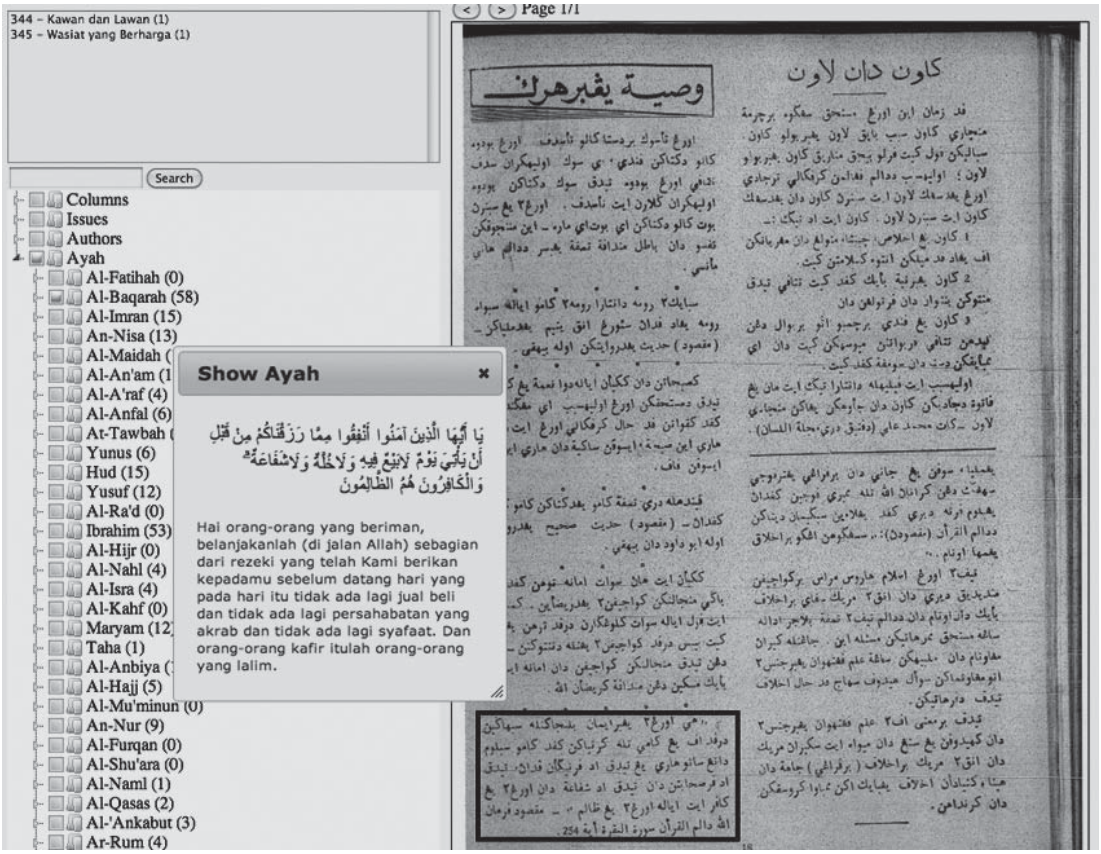

図8　コーランの章句の検索画面

アラビア語とローマ字のそれぞれによる電子版のコーラン[3]を入手し、『カラム』記事データベースにインデックス化した。その際に、『カラム』記事からコーランの参照部分を示すには２つの方法がある。『カラム』の記事中に「コーラン○○の章の○○節」とある場合には、章の名前と文章番号をもとにコーランの章句を検索した。コーランからの引用であるが章の名前や文章番号が明示されていない場合は、『カラム』記事内の文章とマレー語版コーランのテキストを照らしあわせて引用もとの章句を検索した。

(2)PDFビューア

検索結果を示すため、PDF.js[4]に基づいたウェブ・ベースのPDFビューアを構築した。『カラム』のジャウィの記事は横書きで右から左に向けて書かれ、ローマ字翻字の記事は横書きで左から右に向けて書かれる。ローマ字翻字の結果をジャウィ版と左右がほぼ対称になるようにレイアウトして出力し、テキストをローマ字版から抽出して、単純にそれと左右対称の場所が該当するジャウィの記述がある場所だと想定して領域を指定した。

この結果はおおむね合致していると評価できるが、図７に示したように、領域分割してみるとジャウィ版とローマ字版では厳密に左右対称になっておらず、検索結果を示したときに細かい部分で領域がずれるという問題が生じる。なお、これはローマ字版を作成するときに領域を独自に指定するなどの方法で対応することが可能である。

(3)コーランの章句の参照

利用者が『カラム』のジャウィ版の記事を指定すると、それに該当するローマ字記事をもとにコーランの記事が検索され、検索結果がコーランの章句のアラビア語とマレー語訳としてジャウィ版の記事に対して示される（図８[5]。図８では右側にジャウィ版『カラム』記事、左側にコーランの各章のメニューが記されている。検索結果は、『カラム』記事のコーランの章句の引用箇所が四角で囲ってハイライトされるとともに、コーランの内容がアラビア語とマレー語でポップアップ画面で表示される。

3 http://www.qurandatabase.org/
4 http://mozilla.github.com/pdf.js/
5 暫定的なシステムは以下のアドレスで暫定的に公開されている。http://gaia.net.cias.kyoto-u.ac.jp/qalam/ このURLは予告なく変更することがある。

むすび

本稿では、利用者が意味的な注釈に基づいたアプローチによって『カラム』の文脈を理解するのを助けるアプローチを示した。本稿で示されたプロトタイプは、利用者が基礎的なライブラリ検索を越えて『カラム』記事の中のコーランへの言及を捜すことを可能にする。

ただし、その仕事はまだ進行中で、新しい注釈を手動で含める方法はまだない。また、オントロジーへの結び付け、とりわけ［Dukes *et al.* 2011］で示されているイスラム的文脈への結びつけは今後の課題である。

参考文献

Ahmad, Fatimah, Mohammed Yusoff, and Tengku MT Sembok. 1996. “Experiments with a stemming algorithm for Malay words.” *Journal of the American Society for Information Science*. 47(12):909-918.

Amin, Adnan. 1998. “Off-line Arabic character recognition: the state of the art.” *Pattern recognition*. 31(5):517-530.

Beagrie, Neil. 2003. *National digital preservation initiatives*. Council on Library and Information Resources.

Berners-Lee, Tim, James Hendler, and Ora Lassila. 2001. “The semantic web.” *Scientific American*. 284(5):28-37.

Bourdon Julien and Shibayama Mamoru. 2012. “Knowledge Creation in Area Studies: a Semantic-Based Approach.” Culture and Computing 2012. LNCS Transactions on Edutainment IX (251-258), Springer-Verlag.

Broshi, Magen. 2004. “The Dead Sea Scrolls, the sciences and new technologies.” *Dead Sea Discoveries*. pp.133-142.

Budi, Indra, and Stephane Bressan. 2007. “Application of association rules mining to Named Entity Recognition and co-reference resolution for the Indonesian language.” *International Journal of Business Intelligence and Data Mining*. 2(4):426-446.

Dukes, Kais, Eric Atwell, and Nizar Habash. 2011. “Supervised collaboration for syntactic annotation of Quranic Arabic.” *Language Resources and Evaluation*. pp:1-30.

Ghani, Roslan Abdul, Mohamad Shanudin Zakaria, and Khairuddin Omar. 2009. “Jawi-Malay Transliteration.” Electrical Engineering and Informatics, International Conference on. Vol. 1. IEEE, 2009.

Kakali, Constantia, et al. 2007. “Integrating Dublin Core metadata for cultural heritage collections using ontologies.” International Conference on Dublin Core and Metadata Applications.

Lee, John K., and Brendan Calandra. 2004. “Can Embedded Annotations Help High School Students Perform Problem Solving Tasks Using A Web-Based Historical Document?.” *Journal of Research on Technology in Education*. 37:65-84.

M Zeki, Ahmed, Mohamad S Zakaria, and Choong Yeun Liong. 2007. “Isolation of Dots for Arabic OCR using Voronoi Diagrams.” Proceedings of the International Conference on Electrical Engineering and Informatics, 2007.

McGuinness, Deborah L., and Frank Van Harmelen. 2004. “OWL web ontology language overview”. *W3C recommendation 10*. 2004-03:10.

Omar, Khairuddin, et al. 2012. “"Skew Detection and Correction of Jawi Images Using Gradient Direction.” *Jurnal Teknologi*. 37:117-126.

Samat, Talib. 2002. *Ahmad Lutfi: Penulis, Penerbit, dan Pendakwah*. Dewan Bahasa dan Pustaka.

Uren, Victoria, et al. 2006. “Semantic annotation for knowledge management: Requirements and a survey of the state of the art.” *Web Semantics: science, services and agents on the World Wide Web 4.1*. pp.14-28.

Van der Putten, Jan, 2010. “Negotiating the Great Depression: The rise of popular culture and consumerism in early-1930s Malaya.” *Journal of Southeast Asian Studies*. 41(1):21-45.

Yamamoto, Hiroyuki. 2009 “The Jawi publication network and ideas of political communities among the Malay-speaking Muslims of the 1950s Muslim networks and movements in Asia.” *The Journal of Sophia Asian Studies*. 27:51-64.

Zahidah, Z., A. Noorhidawati, and A. N. Zainab. 2011. “Exploring the Needs of Malay Manuscript Studies Community for an E-Learning Platform.” *Malaysian Journal of Library & Information Science*. 16(3):31-47.